

INTERNATIONALIZING 360-DEGREE FEEDBACK: ARE SUBORDINATE RATINGS COMPARABLE?

Treena L. Gillespie

DePaul University

ABSTRACT: This study addressed whether 360-degree feedback ratings made by subordinates are equivalent across national cultures (Great Britain, Hong Kong, Japan, U.S.). Employees' ratings of their leaders were collected via a company's existing 360-degree feedback survey. A two-factor baseline model was developed using exploratory factor analysis and tested across cultures utilizing confirmatory factor analysis techniques. Results revealed differences in factor form, suggesting that the constructs underlying the survey and their relationship to the survey items differed across cultures. These results emphasize the need for multinational companies to use caution when transporting human resources practices, such as 360-degree feedback, to international locations.

KEY WORDS: multisource; 360; measurement equivalence; culture.

Many organizations are experiencing the opportunities yielded by international expansion. However, global growth surfaces critical decision points for multinational companies as they consider how to develop their employees across country borders. Cultural differences influence management approaches and employee expectations, making standard U.S.-based practices difficult to transplant in non-U.S. locations (Cascio & Bailey, 1995). Organizations not considering these differences in their people-management approaches could be missing opportunities to improve organizational effectiveness. Thus, companies face the dilemma of whether to institute standardized human resource (HR) practices globally or to substitute parallel processes for each country that reflect local

The author would like to express appreciation to Alice Stuhlmacher for her valuable comments on earlier versions of this article. This research was conducted as part of the author's doctoral dissertation. Portions of this article were presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Ontario, Canada, April 2002.

Address correspondence to Treena L. Gillespie, Department of Management, California State University-Fullerton, P.O. Box 6848, Fullerton, CA 92834-6848. E-mail: tgillespie@fullerton.edu.

cultural values. One popular HR practice that poses challenges for global use is 360-degree feedback, a tool that allows multiple sources, including subordinates, to provide leaders with feedback geared toward performance development or improvement.

During the past few decades, 360-degree feedback (also known as multisource feedback) emerged and has gained momentum in the performance management literature. Coupled with its gain in popularity has been its expansion throughout companies, its use transcending national boundaries in response to a growing global marketplace. Research on the effectiveness of using 360-degree feedback in international locales has lagged the pace that companies are implementing such systems. Organizations have little information about the projected effectiveness of their U.S.-developed systems in international offices. To date, no research has focused on whether performance ratings gathered from international locations through 360 processes have measurement equivalence, or mean the same thing. Thus, practitioners have limited guidance on how the results from an "internationalized" 360 process can be best utilized. This study addresses this research gap by examining the results of one company's 360-degree feedback system across different national cultures.

EXPANDING 360-DEGREE FEEDBACK GLOBALLY

Augmenting the traditional practice of top-down supervisory performance appraisal, 360-degree feedback entails having multiple sources (e.g., subordinates, customers, peers, supervisor, self) provide performance feedback to a leader in areas important for organizational success (Bracken & Timmreck, 1999; Tornow, 1993). Thus, the leader is offered a multi-dimensional view of his or her performance that reveals discrepancies, consistencies, strengths, and developmental areas. These data provide the targeted leader with a framework for improving behavior and increasing effectiveness.

Currently, many organizations, including a majority of Fortune 500 companies, employ 360-degree feedback, spending millions of dollars each year on the process (Cheung, 1999; Church, 1995; Hazucha, Hezlett, & Schneider, 1993; London & Beatty, 1993; Yammarino & Atwater, 1997). Practical enthusiasm for the tool has been supported by research substantiating its effectiveness. Documented improvements include increased performance (e.g., Reilly, Smither, & Vasilopoulos, 1996; Smither, London, Vasilopoulos, Reilly, Millsap, & Salvemini, 1995), more positive perceptions by subordinates (Atwater, Roush, & Fischthal, 1995; Bernardin, Hagan, Ross, & Kane, 1995; Hegarty, 1974; Reilly et al., 1996; Smither et al., 1995), and an increase in leadership skills (Hazucha et al., 1993). Although the effects of 360-degree feedback are still being

examined, initial perceptions of effectiveness have helped 360-degree feedback gain a foothold in industry.

Once considered a means for developing managerial behaviors, the process has evolved, rightly or wrongly, to include both the individual and the organizational levels, providing data that the company can use in its strategic efforts. Results of 360-degree feedback are used in conjunction with HR systems, such as staffing, succession planning, compensation, performance management and even, in some cases, downsizing (Bracken & Timmreck, 1999). Thus, the quality of the data, while important when used for employee development, has substantial impact when used for administrative purposes.

To the extent that 360-degree feedback is perceived as adding value in terms of managerial and organizational effectiveness, multi-national companies (MNCs) are likely to expand the process globally. Perhaps the easiest and most cost-effective method of implementing 360-degree feedback for international locations simply is to extend the U.S.-based process. The evidence of success in domestic sites may lead business leaders to assume that the effectiveness of multi-rater feedback transcends national boundaries. However, as previously mentioned, there is a paucity of research examining whether this common practice of "internationalizing" 360-degree feedback processes is indeed a sound business practice. Is the survey relevant for other cultures and do the resulting feedback ratings mean the same thing, regardless of culture?

Cultures, commonly defined by country of residence (e.g., Hofstede, 1983; House, Wright, & Aditya, 1997), differ in their shared experiences and norms that govern life within the collective (House et al., 1997; Javidan & House, 2001). Beliefs, or actual practices, as well as values, or preferred practices, vary from culture to culture (Javidan & House, 2001). For example, two strongly supported cultural values that are especially relevant to leadership and 360-degree feedback processes, are Power Distance and Individualism/Collectivism (Hofstede, 1980). Power Distance refers to the degree of centralization of power and autocratic leadership in the nation. Individualism/Collectivism describes how much the culture is defined from the group perspective. A country's position on these values (higher versus lower) tends to define the perceptions shared within the collective and serves to distinguish societies from each other (Javidan & House, 2001).

This past work in cross-cultural research suggests that a U.S.-developed conceptualization of leader effectiveness, as defined by a company's 360-degree feedback survey, may not match the perceptions of effective leadership in countries with different levels of cultural values. Although some aspects of the cultural norms may be similar and some leadership skills may transcend cultures (e.g., Robie, Johnson, Nilsen, & Hazucha, 2001), it is also likely these can differ between the cultures

(Javidan & House, 2001). It follows that a survey founded on a U.S.-based conceptualization of leader effectiveness may be viewed by other cultures as irrelevant to their values, confusing, or simply incomplete by omitting some of the values they hold (Peterson & Hunt, 1997). Aggregating or using these data across cultures for developmental or administrative purposes could lead an organization toward erroneous conclusions. Therefore, it is important to look closer at what the data collected from a 360-degree feedback survey represent and whether they are comparable across cultures.

MEASUREMENT EQUIVALENCE

Traditionally, cross-cultural differences have been investigated either by comparing mean scores or examining correlations on key variables between two or more groups (Bass, 1990; House et al., 1997). However, interpreting group differences without first establishing the equivalence of each group's data results in findings that may not be substantive or accurate (Aycan, Kanungo, & Sinha, 1999; Bond & Smith, 1996; Bontempo, 1993; Drasgow & Kanfer, 1985; Greenfield, 1997; Van de Vijver & Leung, 1997; Vandenberg & Lance, 2000). Thus, the use of mean differences may result in misguided or inaccurate conclusions if the means are derived from data that are non-equivalent across groups (Riordan & Vandenberg, 1994). When comparing or aggregating data from HR instruments such as 360-degree feedback, drawing inaccurate conclusions may have deleterious effects for both individual development and organizational effectiveness.

Equivalence is defined and embodied in research a number of different ways (see Adler, 1983 for a review). Historically, ensuring survey or test equivalence across cultures entailed pretesting the measure, translating it, and having a bilingual expert translate it back to the original language to ensure all meanings were preserved (e.g., Kanungo & Wright, 1983; Spencer-Oatey, 1997), making changes if needed (Brislin, 1980). But, translation is not enough to establish equivalence (Bontempo, 1993; Riordan & Vandenberg, 1994). People from varied environments are likely to use different categories of thinking, or constructs (Hofstede, 1992) due to differences in cultural socialization (Smith, Tisak, Bauman, & Green, 1991). Even if translations are near perfect, these differences in constructs remain, so that items may be tapping into areas that are unfamiliar to some cultures (Berry, Poortinga, Segall, & Dasen, 1992; Matsumoto, 1994; Smith et al., 1991).

In general, scores on identical instruments are considered equivalent when they have the same psychological meaning in different contexts or cultures (Berry et al., 1992). In this study, equivalence refers to whether or not the 360-degree feedback measure relates to a set of underlying

constructs in the same way and to the same extent in multiple groups (Cole & Maxwell, 1985). Specifically, this study defines equivalence at several levels, each providing specific information about what is and is not considered equivalent (Bollen, 1989; Mullen, 1995; Singh, 1995). First, equivalence is tested at the construct level, ensuring that the same constructs exist and relate to each other in the same way across cultures. Beyond this first examination, the second step for assessing equivalence is examining the equality of the relationships between survey items and constructs across each nation. Finally, equivalence is defined in terms of the invariance of measurement error across cultures. This definition of equivalence allows scrutiny of its various aspects and provides specific information as to how the data collected from different cultures compare.

To date, no data have been published detailing the equivalence of ratings made by subordinates of varying national cultures within the context of an existing 360-degree feedback program. Measurement equivalence issues in relation to 360 have focused primarily on the differences between rating sources (e.g., Atwater, Ostroff, Yammarino, & Fleenor, 1998; Cheung, 1999), not on cross-cultural variation for sources. Although methodologically useful to international research, these U.S.-centered studies provide few clues as to the functioning of a U.S.-based survey practice in non-U.S. locations. Thus, the present study offers a unique contribution to theory and practice by further investigating the validity of using 360-degree feedback in a MNCs international sites.

THE PRESENT STUDY

The study examined the measurement equivalence of a 360-degree feedback survey in the U.S. and three international locations of one U.S.-headquartered MNC. Subordinate ratings of their immediate supervisors were used as the basis for analysis.

The four national cultures of interest were Great Britain, Hong Kong, Japan, and the United States. Countries were selected using two criteria: (1) position on cultural values and (2) meaningfulness. Comparatively, Great Britain and the United States tend to have lower Power Distance and higher Individualism, while Hong Kong and Japan maintain higher Power Distance and lower Individualism values (Hofstede, 1980). Also, inclusion of these countries offered practical relevance to the sponsoring organization because these offices were among the largest international operations in the company.

To examine the issue of measurement equivalence, multiple group structure analysis was employed. Specifically, confirmatory factor analysis (CFA) was utilized to examine whether the U.S.-derived measurement model held across Great Britain (GBR), Hong Kong (HKG), and Japan (JPN), as well as an additional U.S. sample. Several research

questions were addressed, with three specific hypotheses tested in relation to the survey data.

The first research question explored whether or not the same constructs existed and related to each other in the same way across cultures. In particular, the groups were tested for a consistent factor pattern. This research question investigated whether the conceptual paths between factors, survey items, and errors were the same across cultures, without testing whether or not the specific values were the same. The hypothesis was represented as:

$$\text{Hypothesis 1: } H_{\text{form}} : \text{form}^{(\text{GBR})} = \text{form}^{(\text{HKG})} = \text{form}^{(\text{JPN})} = \text{form}^{(\text{US})}.$$

A second research question focused on the equivalence in relationships between survey items and constructs across national cultures. This hypothesis defined equivalence in more detail than did Hypothesis 1, establishing the relationship between specific items and constructs or factors, represented by factor loadings (Λ). Specifically, factor loadings were to be examined for invariance across the four cultures. Invariance of factor loadings indicates that the relationship between the construct and each item is the same across cultures (Bollen, 1989). This hypothesis was expressed as:

$$\text{Hypothesis 2: } H_{\Lambda X} = \Lambda_X^{(\text{GBR})} = \Lambda_X^{(\text{HKG})} = \Lambda_X^{(\text{JPN})} = \Lambda_X^{(\text{US})}.$$

The final, more restrictive, research question centered on measurement error. This hypothesis related to the invariance of measurement error (Θ_δ) across all four cultures. Measurement error focuses on both random and systematic error. Therefore, reliability and scalar equivalence were to be assessed (Bollen, 1989; Mullen, 1995). The measurement error hypothesis was represented as:

$$\text{Hypothesis 3: } H_{\Theta_\delta} = \Theta_\delta^{(\text{GBR})} = \Theta_\delta^{(\text{HKG})} = \Theta_\delta^{(\text{JPN})} = \Theta_\delta^{(\text{US})}.$$

Failure to reject these hypotheses would support the equivalence of the respective portion of the measurement model. Rejection of a hypothesis would indicate a lack of equivalence between cultures.

METHOD

Participants

A total of 1,833 leaders of a U.S.-based transportation company participated in the company's 360-degree feedback process. The process included leaders in headquarters and field locations, with field locations representing the operational side of the company and headquarters offering support as well as housing the executive levels of leadership. Jobs, environment, and reporting relationships differed substantially

between field and headquarters. Thus, to enhance generalization and to reduce extraneous factors, only subordinate ratings for leaders in field locations (domestic and international) were included in this study.

Rater demographic data were not collected. The management team considered that collecting demographic data might undermine the anonymity of the process, resulting in lower response rates or inaccurate data. Therefore, sample characteristics, other than work location and function, were not surveyed and are not available for analysis.

Model Development. Overall, 19,256 surveys of leaders were received from subordinates based in U.S. field locations. A sample of 1,000 surveys (5%) was selected randomly to use as the basis for developing an initial model to test across cultures.

Cross-Cultural Comparisons. For cross-cultural model comparisons, direct report ratings were selected randomly from the U.S. ($N = 18,256$, excluding the sample used for model development), GBR ($N = 214$), HKG ($N = 459$), and JPN ($N = 742$) surveys. Disparate sample sizes were anticipated to complicate the types of analyses planned. Therefore, 214 direct report surveys were selected randomly from each country.

Measures

A team (including internal researchers, senior management, employees from different functional areas, and an external consulting firm) constructed the 360-degree feedback survey. The custom survey was designed to address the company's success factors, or areas that defined leadership effectiveness for the organization. These success factors were rationally selected on the basis of their clarity, importance, and how observable they are on the job.

After selecting the dimensions as the survey framework, an external vendor furnished a pool of items that were theoretically related to the dimensions. The team added items to this pool. The entire pool of items was then distributed to a sample of employees who provided their comments on the clarity, importance, and observability of each item. Items higher on these criteria were retained in the survey, while less clear, important, or observable items were excluded. The final instrument contained 40 items assessing aspects of leadership behaviors the company was striving to encourage, as well as open-ended questions inviting comments. Respondents indicated the extent to which their leader engaged in a behavior using a 5-point Likert-type scale, with 1 indicating "Not at All" to 5, indicating "To a Very Great Extent." In addition,

respondents could choose to rate "Not Applicable" for behaviors that they had not had the opportunity to observe.

The survey was translated into multiple languages (Japanese, Mandarin, Spanish) using an external language-consulting firm, with back-translation and modifications made by internal experts. English was retained on the translated surveys so that employees could choose which language they felt most comfortable using.

Procedure

The purpose of the 360-degree feedback process was to help leaders develop effective behaviors by learning their strengths and their developmental areas. Administrative decisions (e.g., promotions, pay increases) were not tied to the survey results.

A nomination package was sent by company mail to each feedback recipient at his or her work location. The package included an introductory letter, information about the process, a nomination form, and a pre-addressed envelope for returning the completed nomination form. On the nomination form, the participating leader was asked to identify his or her supervisor, peers, direct reports, and internal customers. Therefore, feedback recipients were chosen by leaders rather than assigned. To mitigate leaders' attempts to "stack the deck" in their favor, they were instructed to submit their employee roster as part of their nomination form. Completed nomination forms were returned to the external vendor.

Survey packages were sent to all individuals who were listed as a feedback provider on one or more nomination forms. Each package contained information about the process, instructions for completing the surveys, the survey(s), and a postage-paid envelope for direct return to the external vendor. Numerous actions were taken to ensure a reasonable response rate, including education about the process, reminders through internal news sources and posters, and the establishment of a dedicated e-mail address and voice mailbox for fielding questions and concerns.

Domestic (U.S.) employees returned their surveys directly to the external vendor using a postage-paid envelope. International employees were instructed to seal their survey in an envelope and forward it to a designated point of contact (POC) in their country. The POC then shipped all surveys to the external vendor via a commercial delivery company.

Upon receiving the surveys, the external vendor analyzed the results and developed feedback reports for the feedback recipients. Comments made on the survey were transcribed verbatim and included in the recipient's report. Participants received their reports at workshops, where

trained facilitators assisted them in interpreting their feedback and beginning the action planning process.

RESULTS

Model Development

The 360-degree feedback instrument consisted of rationally derived factors and items. In the absence of strong theory or conclusive results from previous studies, the initial baseline model was derived from a principal components exploratory factor analysis (EFA) using a sample of U.S.-based subordinates ($N = 1,000$).

All 40 items were included in the EFA. The initial solution revealed two factors with eigenvalues greater than 1.0. Factor 1 (initial eigenvalue = 28.69) accounted for 71.71 percent of the variance and Factor 2 (initial eigenvalue = 1.03) accounted for 2.57 percent; thus, when combined, 74.28 percent of the variance was accounted for by these two factors. The scree plot indicated a clear break after Factor 1. However, the second factor was retained in anticipation that rotation would allow examination of a clearer structure and the possibility to compare the results to theoretical underpinnings.

After orthogonal rotation, Factor 1 (rotated eigenvalue = 16.15) accounted for 40.37 percent of the variance found in the original variables; Factor 2 (rotated eigenvalue = 13.56) accounted for 33.91 percent. The analysis was then conducted using oblique rotation. The survey items originally were developed to measure leader effectiveness, so it was possible that the factors were related. After the oblimin rotation, the new eigenvalues for Factors 1 and 2 were 28.27 and 1.84, respectively. Examination of the factor correlation matrix revealed that Factors 1 and 2 were relatively independent. Their correlation was .15, indicating that only 2.2 percent of the variance overlapped among the factors. Therefore, the final solution from the orthogonal rotation was used for the remaining analyses.

Factor loadings were then examined to identify which items were the most integral components of each factor. Comrey (1973) proposed a guideline for interpreting loadings, with the criterion being the percentage of overlapping variance between the item and the factor. For example, if an item has 50 percent of its variance explained by a factor, its loading on that factor would be .71. Loadings of .63 (40 percent overlapping variance) are considered very good, .55 (30 percent overlapping variance) good, and .45 (20 percent overlapping variance) fair; smaller loadings are considered poor (Comrey, 1973). A factor loading of .71 was chosen as the criterion for retaining items for further analysis. The larger number of items in the survey allowed this more restrictive

Table 1
Survey Item Descriptive Statistics

Survey Item	N	M	SD
3. Treats others fairly, with respect and without prejudice	979	3.98	1.21
6. Listens attentively to the comments and ideas of others	983	3.98	1.16
7. Trusts and encourages others to make decisions	933	3.95	1.15
9. Translates long range goals into concrete plans for action	834	3.66	1.18
10. Demonstrates expertise in his/her functional area of the business	960	3.96	1.19
14. Insists on safe work practices regardless of other constraints	828	4.14	1.00
16. Recognizes, appreciates and utilizes the unique knowledge, skills and abilities of others	952	3.85	1.23
20. Holds others accountable for their decisions and actions	906	3.80	1.15
21. Effectively balances costs and benefits when making decisions	753	3.83	1.13
22. Demonstrates long-term thinking that goes beyond the current plan year	737	3.72	1.21
26. Is readily accessible to discuss problems or issues	981	3.81	1.30
27. Adapts easily to new ideas or suggestions	902	3.64	1.22
29. Is ethical and trustworthy	965	4.10	1.23
30. Supports the efforts of others to innovate and take risks	862	3.73	1.23
31. Develops effective and realistic plans	834	3.74	1.15
32. Leads by example	926	3.69	1.30
33. Seeks feedback from direct reports and peers	877	3.69	1.29
34. Recognizes others for their successes	963	3.88	1.25
35. Works well with other departments	793	4.08	1.08
36. Is positive and supportive	978	3.98	1.25

Note. Respondents indicated the extent to which the leader engaged in a behavior using a 5-point scale (1 = not at all, 5 = to a very great extent).

choice. In addition, some items had relatively high levels of shared variance with both factors. Focusing on items that clearly loaded on only one factor provided a cleaner structure for subsequent analyses. Twenty items were retained for further analyses. The descriptive statistics for these items are exhibited in Table 1.

Factor 1 (Interpersonal; $\alpha = .978$) included thirteen items relating to interpersonal aspects of the leader's job, while Factor 2 (Task; $\alpha = .952$) included seven task-oriented items. The factor loadings of the items on the factors are presented in Tables 2 and 3, respectively. These factors conceptually are similar to the established leadership dimensions of Consideration and Initiating Structure. Historically, these two categories of behavior have been cited as describing how subordinates perceive their supervisor's behavior (Fleishman, 1953; Halpin & Winer, 1957; Hemphill & Coons, 1957) and are relatively independent from each other (Yukl, 1998), consistent with the findings of the EFA. This resulting baseline model was subsequently tested across four cultures (GBR, HKG, JPN, U.S.) using CFA techniques.

Table 2
Loadings on Factor 1 (rotated eigenvalue = 16.15; α = .978) for Principal Components Exploratory Factor Analysis

Survey Item	Loading
36. Is positive and supportive	.84
34. Recognizes others for their successes	.82
3. Treats others fairly, with respect and without prejudice	.76
16. Recognizes, appreciates and utilizes the unique knowledge, skills and abilities of others	.76
32. Leads by example	.73
35. Works well with other departments	.73
27. Adapts easily to new ideas or suggestions	.73
30. Supports the efforts of others to innovate and take risks	.73
33. Seeks feedback from direct reports and peers.	.73
26. Is readily accessible to discuss problems or issues	.72
6. Listens attentively to the comments and ideas of others	.71
7. Trusts and encourages others to make decisions	.71
29. Is ethical and trustworthy	.71

Table 3
Loadings on Factor 2 (rotated eigenvalue = 13.56; α = .952) for Principal Components Exploratory Factor Analysis

Survey Item	Loading
21. Effectively balances costs and benefits when making decisions	.80
22. Demonstrates long-term thinking that goes beyond the current plan year	.75
14. Insists on safe work practices regardless of other constraints	.72
10. Demonstrates expertise in his/her functional area of the business	.72
9. Translates long range goals into concrete plans for action	.71
20. Holds others accountable for their decisions and actions	.71
31. Develops effective and realistic plans	.71

Model Fit Across Cultures

Multiple group structure analysis was conducted using LISREL 8.30 software (Jöreskog & Sörbom, 1999). This procedure allowed hypotheses about the baseline model to be tested across samples using CFA.

Given LISREL's constraints in processing missing data, the data files were examined to determine the extent of blank items or items marked "not applicable (NA)." There were 4,280 possible data points for each sample (i.e., 20 items for each of the 214 respondents). Out of this total, 12 percent of the U.S. sample's data points were missing or marked "NA," followed by HKG (11 percent), JPN (10 percent), and GBR (8 percent). However, the missing data tended to cluster by respondents differently across samples. More of the HKG respondents ($N = 108$) left items blank or marked "NA," followed by GBR ($N = 91$), the U.S.

($N = 89$), and JPN ($N = 83$). Thus, across samples, between 39 and 50 percent of each sample responded with "NA" or left one or more items blank.

Missing data also tended to cluster by item. For all samples, items 21 ("Effectively balances costs and benefits when making decisions") and 22 ("Demonstrates long-term thinking that goes beyond the current plan year") drew the largest number of missing data points. Between 19 and 29 percent of each sample had no data for item 21; for item 22, between 18 and 28 percent of the data were missing. Although the amount of missing data for these items is substantial, the items were retained for further analyses, given their high loadings on Factor 2 in the EFA.

Missing data were estimated using the mean for each group, a conservative method of data replacement. However, mean replacement could decrease variability, potentially lowering the values in the correlation matrix. Given that each item had at least 70 percent of respondents answering, it was unlikely that the estimations would substantially impact the results. However, as a precaution, the correlation matrices reflecting missing data versus mean replacement were compared for each culture. No differences existed, providing support for using mean replacement as a method for estimating the missing data.

Before testing the hypotheses, the ordinal-type raw data were converted to polychoric correlation matrices and asymptotic covariance matrices. Measurement models were constructed for each sample, defined by the parameter matrices (Λ , Θ , and Φ). Estimates were fitted according to the Weighted Least Squares (F_{WLS}) fitting function, a suitable function for estimating the model with ordinal data (Bollen, 1989; Jöreskog & Sörbom, 1993; Mullen, 1995). Analyses were conducted in the same order as the hypotheses.

The first research question called for establishing whether or not the same constructs existed and related to each other the same way across cultures. The hypothesis for this question was focused on testing the invariance of factor form across cultures, or $H_{\text{form}} : \text{form}^{(\text{GBR})} = \text{form}^{(\text{HKG})} = \text{form}^{(\text{JPN})} = \text{form}^{(\text{US})}$. Each culture was specified as a separate group in the multiple group structure analysis.

Initially, the model was rejected ($\chi^2 = 4832.84$, $df = 676$, $p < .000$), indicating that a lack of equivalence existed between cultures. However, given the number of cases in this study, additional fit indices were used that were less reliant on sample size.

Global goodness-of-fit statistics were reviewed to determine how well the data from each of the cultures fit the baseline model. These indices included the root mean square error of approximation (RMSEA) and comparative fit index (CFI; Bentler, 1990). Analysis of H_{form} revealed an RMSEA value of .17, signifying that the error of approximation was too large for assumptions of close fit. The CFI for H_{form} was .99, suggesting

that the baseline model from the EFA fit the data better than a model with no relationships among its variables.

Group fit statistics were evaluated as well. Specifically, the goodness-of-fit index (GFI) and the standardized root mean square residual (SRMSR; Jöreskog & Sörbom, 1993) values were consulted. The GBR, HKG, JPN, and U.S. samples had GFI values of .97, .94, .95, and .95, respectively. Although high, these values are lower than what is often found with models of good fit. In addition, Bollen (1989) suggested that GFI values could remain high even when the models are seriously misspecified. Thus, the GFI offers one piece of information, albeit inconclusive, in the larger picture of model fit. The SRMSR values for the groups ranged from .16 to .23, beyond the typical criterion of .10 designating acceptable fit. Hong Kong and Great Britain samples had the largest values (.23 and .22, respectively) with Japan and the U.S. having the smallest value of average residuals (.19 and .16, respectively).

Taken together, the indices of fit, along with the χ^2 test results indicated the need to reject the first hypothesis, H_{form} . The rejection of H_{form} suggests that the factor pattern was not the same across all groups. Specifically, these results imply that the data from the different cultures define a different conceptual model than that identified in the EFA using a U.S. sample. Differences could have existed in the number of factors, whether a relationship existed between them, which observed items the factors influenced, or the relationships among measurement error.

Without supporting evidence that the relationships in the model were conceptualized the same across cultures, it did not make sense to test the invariance of these relationships. Therefore, the results from H_{form} precluded testing Hypothesis 2 or Hypothesis 3. The H_{form} is the least demanding test, so it was not appropriate to advance to the more restrictive tests without support for H_{form} (Bollen, 1989).

Exploratory Analyses

Rejection of H_{form} provides little indication as to what aspects of the baseline model failed to be replicated in other cultures. No clues were available as to which part of the model form was unsubstantiated. Thus, exploratory analyses were conducted to obtain possible signs for why the first hypothesis might have been rejected.

Each construct was tested individually by culture. Modifications were made to the model for each construct until a fit was obtained. The models for each construct were then combined into the complete model and fit was assessed. Again, modifications were made until the comprehensive model indicated reasonable fit to the data. Although this method, without proper cross-validation, may capitalize on chance or the characteristics of the specific samples, it was pursued in anticipation of obtaining clues about model differences on an exploratory basis.

Great Britain. The Interpersonal construct (Factor 1) fit marginally in relation to the base model's hypothesized items ($\chi^2 = 140.24$, $df = 65$, $p < .000$; RMSEA = .074; CFI = .97; GFI = .98; SRMSR = .12). However, specifying correlated error terms improved model fit ($\chi^2 = 75.74$, $df = 56$, $p < .041$; RMSEA = .041; CFI = .99; GFI = .99; SRMSR = .093). The baseline model for the Task factor (Factor 2) fit the data reasonably well ($\chi^2 = 29.67$, $df = 14$, $p < .009$; RMSEA = .072; CFI = .97; GFI = .99; SRMSR = .063). However, review of the R^2 estimates revealed that the factor was accounting for very little variance in items 14 ("Insists on safe work practices") and 20 ("Holds others accountable"). Deleting these paths and allowing correlated error covariance improved the fit ($\chi^2 = 4.41$, $df = 4$, $p < .35$; RMSEA = .022; CFI = 1.00; GFI = 1.00; SRMSR = .024).

Combined, the GBR model showed poor fit ($\chi^2 = 315.53$, $df = 124$, $p < .000$; RMSEA = .085; CFI = .97; GFI = .97; SRMSR = .17). Modifications were made after inspecting the R^2 values, residuals, and modification indices. Most notably, correlating error variances contributed to a better fitting model as well as the exclusion of items 21 ("Effectively balances costs and benefits when making decisions") and 33 ("Seeks feedback from direct reports and peers"). Finally, item 3 ("Treats others fairly") was predicted by both factors. In sum, the respecified model allowed error terms to correlate and eliminated paths to four items (items 14, 20, 21, and 33). In addition, one item was explained by both constructs. These changes yielded a model that fit the data ($\chi^2 = 104.76$, $df = 86$, $p < .083$; RMSEA = .032; CFI = 1.00; GFI = .99; SRMSR = .10).

Hong Kong. The baseline model for Factor 1 (Interpersonal) was just beyond reasonable fit ($\chi^2 = 161.35$, $df = 65$, $p < .000$; RMSEA = .083; CFI = .99; GFI = .98; SRMSR = .12). Based on an analysis of the residuals, explained variance, and the modification indices, certain error variances were allowed to correlate. The addition of correlated errors enhanced the fit of the model to the data ($\chi^2 = 85.70$, $df = 59$, $p < .013$; RMSEA = .046; CFI = 1.00; GFI = .99; SRMSR = .093). The baseline model for Factor 2 (Task) was evaluated for fit against the HKG data. The model fit with no alterations ($\chi^2 = 22.19$, $df = 14$, $p < .075$; RMSEA = .052; CFI = .99; GFI = .99; SRMSR = .051).

In testing the entire model, the above changes were made; however, the complete model did not show a good fit with the data ($\chi^2 = 1546.60$, $df = 163$, $p < .000$; RMSEA = .20; CFI = .98; GFI = .98; SRMSR = .24). Numerous alterations to the model were made, with the continued review of item loadings, residuals, and modification indices. Both factors predicted items 3 ("Treats others fairly"), 29 ("Is ethical and trustworthy"), 34 ("Recognizes others for their successes"), and 22 ("Demonstrates long-term thinking that goes beyond the current plan year"). The factors

failed to predict several items, so these paths were deleted. The deleted paths included items 14 ("Insists on safe work practices"), 20 ("Holds others accountable for their decisions and actions"), 30 ("Supports the efforts of others to innovate and take risks"), and 35 ("Works well with other departments"). Error terms were allowed to correlate when suggested by the data and were reviewed for logical relevance. These revisions produced a model that fit the data ($\chi^2 = 127.12$, $df = 86$, $p < .003$; RMSEA = .047; CFI = 1.00; GFI = .99; SRMSR = .090).

Japan. The hypothesized relationships for Factor 1 were isolated to determine fit of the data to this construct. Initially, the data did not fit the model ($\chi^2 = 140.90$, $df = 65$, $p < .000$; RMSEA = .074; CFI = .99; GFI = .99; SRMSR = .14). However, review of the loadings revealed that Factor 1 strongly predicted the observed variables. As with the other samples, one of the key aspects to altering the model to better fit the data involved correlating the error variances for the observed items. The addition of correlated errors resulted in a model that fit the data reasonably well ($\chi^2 = 73.44$, $df = 49$, $p < .013$; RMSEA = .048; CFI = 1.00; GFI = .99; SRMSR = .11). Contrary to the findings for Factor 1, the Japan data readily fit Factor 2. Initially, the fit was marginal ($\chi^2 = 37.45$, $df = 14$, $p < .001$; RMSEA = .089; CFI = .98; GFI = .99; SRMSR = .073), but improved after allowing error terms to covary between items 14 and 20 ($\chi^2 = 23.49$, $df = 13$, $p < .036$; RMSEA = .062; CFI = .99; GFI = .99; SRMSR = .04).

Interestingly, the model combining the two factors had inadequate fit ($\chi^2 = 606.27$, $df = 145$, $p < .000$; RMSEA = .12; CFI = .98; GFI = .98; SRMSR = .15). Observing the large correlation between the factors (.99), a one-factor model was tested. Loadings, residuals, and modification indices were examined for clues as to meaningful model alterations. A number of items had correlated error terms. In addition, paths to three items (14, 20, 33) were removed. This resulted in a better fitting model, although still above the commonly accepted criteria of good fit ($\chi^2 = 187.94$, $df = 108$, $p < .000$; RMSEA = .059; CFI = .99; GFI = .99; SRMSR = .15). Despite numerous attempts to respecify the model, no improvements in fit were made.

United States. Although the baseline model was developed from a U.S.-based sample, the sample used in LISREL analysis did not fit the model. Primarily, the fit indices reported earlier supported rejection of the model; however, these values suggested a closer fit than that found with the HKG and JPN samples, but similar to the GBR results. Each of the factors fit the U.S. data reasonably well when tested independently. The baseline model for Factor 1 resulted in an acceptable fit ($\chi^2 = 121.23$, $df = 65$, $p < .000$; RMSEA = .064; CFI = .99; GFI = .99; SRMSR = .090).

Similarly, no alterations were needed for Factor 2; the data fit the baseline model very well ($\chi^2 = 10.90$, $df = 14$, $p < .690$; RMSEA = .00; CFI = 1.00; GFI = 1.00; SRMSR = .028).

Because no changes were made to either parts of the model, the initial baseline model was used as the starting point for respecifying an overall model that would adequately fit the data. Changes were made based on residuals, loadings, and modification indices, but an exceptionally fitting model failed to emerge. Error terms were correlated when the data and theory suggested it. In addition, the path to item 14 was fixed given its lack of contribution to the fit of the overall model. Item 9 ("Translates long range goals into concrete plans for action") was predicted by both factors. After the respecifications, the final model showed marginal fit ($\chi^2 = 237.43$, $df = 137$, $p < .000$; RMSEA = .059; CFI = 1.00; GFI = 1.00; SRMSR = .13).

The rejection of H_{form} held no clues as to which portions of the model were not substantiated in the four cultures. The exploratory analyses identified several potential areas that could differ from the proposed model. First, the GBR, HKG, and U.S. samples fit a two-factor model, although the JPN data fit a one-factor model best. Secondly, the relationships between survey items and constructs differed across nations. Item 14, relating to safety, offered little or no value to each country's model; however, the inclusion or deletion of other items varied by culture. Finally, the presence of correlated error terms emerged for all countries, although the patterns differed by country. In addition, the GBR sample exhibited generous amounts of error (.20–.49). Thus, these exploratory analyses point to the difficulty in isolating one area that differed from country to country and emphasize the need for using caution in interpreting survey results across national cultures.

DISCUSSION

This study revealed that employees in international locations of the same multinational company interpreted and responded to the same U.S.-developed survey differently. Despite high-quality translations, raters in four different countries (GBR, HKG, JPN, U.S.) had different conceptualizations of the constructs included on the survey and how these constructs relate to each other. Although there may be many potential explanations for the differences between groups, these results call into question the effectiveness and the assumptions of using subordinate ratings in making decisions about employees. These findings have important implications for those responsible for 360-degree feedback efforts, as well as other HR practices.

First, as supported by previous research (Bontempo, 1993; Riordan & Vandenberg, 1994) and demonstrated in this study, translating the survey is not enough. Practitioners also should consider the relevance of the survey constructs and items for each culture. While pretesting a survey, for example, the measurement equivalence could be evaluated, providing information as to which aspects are consistent across cultures and which may need to be customized. Thus, while striving for consistent HR practices, implementation of a global 360 process may necessitate the balance between consistency in the instrument and the differentiation that makes it more relevant for other national cultures. In concept, the organization may define general areas of importance to business success, but leave it up to the international offices how these areas are translated into specific behaviors.

In addition, these results suggest culture may impact individual leaders' use of their data for developmental purposes. The inclusion of feedback workshops is a key factor in the success of a 360-degree feedback process (Bracken & Timmreck, 1999; Yukl, 1998). During these workshops, leaders are often given comparison baseline data, such as company or division means. In light of the findings from this study, aggregation of data to calculate the baseline may be inaccurate. Specifically, the average rating across a global company would reflect ratings made using different constructs and would not serve as the true mean for the company on that item. To increase the effectiveness of 360-degree feedback, aspects such as report clarity and the quality of participant training are integral (Walker & Smither, 1999). Therefore, it may be more advantageous for a global company to offer country norms for each location, giving participants a purer measure against which they can compare their results. Similarly, historical trending may be misleading if the leader held various international posts while participating in successive 360-degree feedback processes.

Finally, these results have practical ramifications at the company level. The data from 360-degree feedback surveys often are aggregated for organizational development and for administrative decisions related to staffing, succession planning, compensation, performance management, and downsizing (Bracken & Timmreck, 1999; Graddick & Lane, 1998). The findings of this study suggest that aggregating subordinate responses may culminate in misguided or inaccurate conclusions if decision-makers do not consider cultural aspects. The results emphasize the need for organizations to consider multiple data sources when making decisions, not just subordinate ratings. Thus, 360-degree feedback is valuable, but is not sufficient for making complex decisions about human resources within a multi-national company; other sources of organizational data (e.g., productivity measures) should be considered.

Aside from the practical implications, the results also have theoretical importance. Although ratings were made within the same

organization, they covaried with national culture. In particular, non-predicted items varied by group. One item measuring how well the supervisor seeks feedback from direct reports and peers was not predicted from the Asian or Great Britain models. Assumptions of cultural precursors to the exclusion of this item are tenuous. However, it is possible that societal position on Power Distance and Individualism/Collectivism could have been operative in its exclusion. Countries high in Power Distance and Collectivism tend to expect autocratic and parental-type leadership (Hofstede, 1980, 1983). They expect their supervisors to become involved in their personal lives (Smith, Dugan, & Trompenaars, 1996) and to provide them with instruction at work, rather than have their supervisor consult with them (Hofstede, 1992). Thus, if HKG and JPN are high in Power Distance and more collectivistic as past research suggests, these findings are consistent with expectations. However, support must be extrapolated from the exploratory analyses, requiring extensive speculation. For example, the Western countries appeared to differ from each other, as did the Asian countries, inconsistent with predictions of the cultural framework.

Additionally, the differences between international locations in this study of one organization lend support for the strength of national culture over organizational culture. For example, Laurent (1992) established that, within one MNC, the management assumptions held by members of cultural subgroups diverged. In the present study, the same result was found, with the meaning of the survey deviating from the structure espoused by senior management. However, the organizational culture in this company was relatively new, with the 360-degree feedback process expected to assist in facilitating culture change. It is possible that the new direction begun at headquarters had not yet seeped into international locales. The employees in the distant offices may have had less opportunity to become accustomed to the behaviors assessed by the survey or to acknowledge and fully understand the new corporate direction. One possibility for future research is investigating the contributions of corporate culture and national culture in feedback ratings.

Several aspects of the survey process could have impacted the results obtained in this study. First, the survey consisted of behavioral-based items rather than trait-based statements. Less individualistic cultures tend to focus on traits such as loyalty, morality, and integrity when evaluating performance (Schneider, 1992). Therefore, different results might be found if investigating the measurement equivalence of a trait-based instrument across international samples. Further, the EFA, using a U.S. sample, identified two factors consistent with U.S. conceptualizations of leadership. Use of this emic approach could have resulted in omission of constructs that exist in other cultures (Bass, 1990; Peterson & Hunt, 1997). Differences in constructs used by each culture could

have led to dissimilar patterns of correlated error variances. However, from a practical perspective, this U.S.-based organization defined leadership within a U.S. framework. Thus, the survey was designed as somewhat prescriptive in that it outlined the behaviors that were expected of all leaders, regardless of location.

A second consideration is the effect of demographics on the results. Demographic data were not collected for this organization. However, it is possible that the findings could be influenced by sub-cultural group membership. For example, Ghorpade, Hattrup, and Lackritz (1999) established that personality measures were not equivalent between Indian men and Indian women. Their finding emphasizes that heterogeneity exists among cultures, with subcultures possibly exerting substantial effects. In addition, knowing respondent nationality could have been helpful in understanding the results. Although within-country variability is likely smaller than between-country variability (Markus & Kitayama, 1991; Triandis, 1993), intrasample diversity could play a substantial role in the findings. This is especially relevant if expatriation is prevalent or if there has been a recent spike in immigration. Recent expatriates or immigrants would have had less time to assimilate into the host country's culture. Thus, the effects of demographics on the measurement equivalence of feedback results should be examined more thoroughly.

Finally, this study examined subordinate ratings, only one of several sources of raters available in most 360-degree feedback programs. The goal of delineating measurement equivalence issues involved in the use of human resource systems predicated the decision to use subordinate ratings exclusively. Also, targets tend to pay more attention to the negative ratings provided by their subordinates than data from other sources in a 360-degree feedback program (Brutus, London, & Martineau, 1999). Therefore, subordinate ratings were selected over the other sources as a first step in examining cross-cultural equivalence within multisource feedback systems. However, results may differ if the data from other rating sources are used. Different rater sources have varied relationships with the feedback target and have unique opportunities for observing the individual. Thus, their ratings are often different (Harris & Schaubroeck, 1988; Wohlers & London, 1989), each including unique variance (Mount, Judge, Scullen, Sytsma, & Hezlett, 1998). For example, supervisors of the targeted feedback recipients may provide more similar results as a group than other sources (Mount et al., 1998). Furnham and Stringfield (1994) found no differences between the ratings of Chinese and European managers when rated by their managers. Past research has surfaced the similarity between subordinate and supervisor ratings (Furnham & Stringfield, 1994; Harris & Schaubroeck, 1988; Holzbach, 1978). Future research efforts should include source comparisons coupled

with cultural differences to determine whether or not measurement equivalence is impacted by the interaction between source and culture.

Developing global-level leadership talent is a salient issue for international companies in the future (Cascio & Bailey, 1995). Multinational organizations struggle between implementing standardized processes, focusing on product and service uniformity, developing consistent leadership, and recognizing cultural uniqueness. Often, the easiest and more cost-efficient choice of managing human resources and developing global talent is in implementing U.S.-based processes such as 360-degree feedback. This study suggests that the assumption of a U.S.-developed survey retaining its meaning across cultures may be faulty. No longer can data be taken for granted; we must recognize that the data collected from identical processes may mean different things to different people. Achieving a competitive advantage in the global marketplace is likely to require more focus on how we manage human resources relevant to our employees' needs, and less on conducting business as usual.

REFERENCES

- Adler, N. J. (1983). A typology of management studies involving culture. *Journal of International Business Studies*, 14(3), 29–47.
- Atwater, L. E., Ostroff, C., Yammarino, F. J., & Fleenor, J. W. (1998). Self-other agreement: Does it really matter? *Personnel Psychology*, 51, 577–598.
- Atwater, L., Roush, P., & Fischthal, A. (1995). The influence of upward feedback on self and follower ratings of leadership. *Personnel Psychology*, 48, 35–59.
- Aycan, Z., Kanungo, R. N., & Sinha, J. B. P. (1999). Organizational culture and human resource management practices: The model of culture fit. *Journal of Cross-Cultural Psychology*, 30(4), 501–526.
- Bass, B. M. (1990). *Handbook of leadership: A survey of theory and research*. New York: Free Press.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238–246.
- Bernardin, J. H., Hagan, C., Ross, S., & Kane, J. S. (1995, May). *The effects of a 360-degree appraisal system on managerial performance*. Paper presented at the 10th annual conference of the Society for Industrial and Organizational Psychology, Orlando, FL.
- Berry, J. W., Poortinga, Y. H., Segall, M. H., & Dasen, P. R. (1992). *Cross-cultural psychology: Research and applications*. New York: Cambridge University Press.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.
- Bond, M. H., & Smith, P. B. (1996). Cross-cultural social and organizational psychology. *Annual Review of Psychology*, 47, 205–235.
- Bontempo, R. (1993). Translation fidelity of psychological scales: An item response theory analysis of an individualism-collectivism scale. *Journal of Cross-Cultural Psychology*, 24(2), 149–166.
- Bracken, D. W., & Timmreck, C. W. (1999). Guidelines for multisource feedback when used for decision making. *The Industrial-Organizational Psychologist (TIP)*, 36(4), 64–74.
- Brislin, R. W. (1980). Translation and content analysis of oral and written materials. In H. C. Triandis & W. Lonner (Eds.), *Handbook of cross-cultural psychology: Methodology* (Vol. 2, pp. 389–444). Boston: Allyn and Bacon.
- Brutus, S., London, M., & Martineau, J. (1999). The impact of 360 degree feedback on planning for career development. *Journal of Management Development*, 18, 676–693.

- Cascio, W., & Bailey, E. (1995). International human resource management: The state of research and practice. In O. Shenkar (Ed.), *Global perspectives of human resource management* (pp. 15–36). Englewood Cliffs, NJ: Prentice Hall.
- Cheung, G. W. (1999). Multifaceted conceptions of self-other ratings disagreement. *Personnel Psychology*, 52, 1–36.
- Church, A. H. (1995). First-rate multi-rater feedback. *Training and Development*, 49, 42–43.
- Cole, D. A., & Maxwell, S. E. (1985). Multitrait-multimethod comparisons across populations: A confirmatory factor analytic approach. *Multivariate Behavioral Research*, 20, 389–417.
- Comrey, A. L. (1973). *A first course in factor analysis*. New York: Academic Press.
- Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, 72, 19–29.
- Fleishman, E. A. (1953). The description of supervisory behavior. *Personnel Psychology*, 37, 1–6.
- Furnham, A., & Stringfield, P. (1994). Congruence of self and subordinate ratings of managerial practices as a correlate of supervisor evaluation. *Journal of Occupational and Organizational Psychology*, 67, 57–67.
- Ghorpade, J., Hattrup, K., & Lackritz, J. R. (1999). The use of personality measures in cross-cultural research: A test of three personality scales across two countries. *Journal of Applied Psychology*, 84(5), 670–679.
- Graddick, M. M., & Lane, P. (1998). Evaluating executive performance. In J. W. Smith (Ed.), *Performance appraisal: State-of-the-art in practice* (pp. 370–403). San Francisco: Jossey-Bass.
- Greenfield, P. M. (1997). Culture as process: Empirical methods for cultural psychology. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology: Theory and method* (Vol. 1). Boston: Allyn and Bacon.
- Halpin, A. W., & Winer, B. J. (1957). A factorial study of the leader behavior descriptions. In R. M. Stogdill & A. E. Coons (Eds.), *Leader behavior: Its description and measurement*. Columbus, OH: Ohio State University.
- Harris, M. H., & Schaubroeck, J. (1988). A meta-analysis of self-supervisory, self-peer, and peer-supervisor ratings. *Personnel Psychology*, 41, 43–62.
- Hazucha, J. F., Hezlett, S. A., & Schneider, R. J. (1993). The impact of 360-degree feedback on management skills development. *Human Resource Management*, 32, 325–351.
- Hegarty, W. H. (1974). Using subordinate ratings to elicit behavior change in supervisors. *Journal of Applied Psychology*, 59, 764–766.
- Hemphill, J. K., & Coons, A. E. (1957). Development of the leader behavior description questionnaire. In R. M. Stogdill & A. E. Coons (Eds.), *Leader behavior: Its description and measurement*. Columbus, OH: Ohio State University.
- Hofstede, G. (1980). *Culture's consequences: International differences in work-related values*. Beverly Hills, CA: Sage Publications.
- Hofstede, G. (1983). The cultural relativity of organizational practices and theories. *Journal of International Business Studies*, Fall, 75–89.
- Hofstede, G. (1992). Cultural dimensions in people management: The socialization perspective. In V. Pucik, N. M. Tichy, & C. K. Barnett (Eds.), *Globalizing management: Creating and leading the competitive organization* (pp. 139–158). New York: John Wiley & Sons.
- Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self, and peer ratings. *Journal of Applied Psychology*, 63, 579–588.
- House, R. J., Wright, N. S., & Aditya, R. N. (1997). Cross-cultural research on organizational leadership: A critical analysis and a proposed theory. In P. C. Earley & M. Erez (Eds.), *New perspectives on international industrial/organizational psychology* (pp. 535–625). San Francisco, CA: New Lexington Press.
- Javidan, M., & House, R. J. (2001). Cultural acumen for the global manager: Lessons from project GLOBE. *Organizational Dynamics*, 29(4), 289–305.
- Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago: Scientific Software.
- Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8.30* [Computer software]. Chicago: Scientific Software.
- Kanungo, R. N., & Wright, R. W. (1983). A cross-cultural comparative study of managerial job attitudes. *Journal of International Business Studies*, 14, 115–129.

- Laurent, A. (1992). The cross-cultural puzzle of international human resource management. In V. Pucik, N. M. Tichy, & C. K. Barnett (Eds.), *Globalizing management: Creating and leading the competitive organization* (pp. 174–184). New York: John Wiley & Sons.
- London, M., & Beatty, R. W. (1993). 360-degree feedback as a competitive advantage. *Human Resource Management*, 32(2/3), 353–372.
- Markus, H., & Kitayama, S. (1991). Culture and self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98, 224–253.
- Matsumoto, D. (1994). *Cultural influences on research methods and statistics*. Pacific Grove, CA: Brooks/Cole.
- Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R., & Hezlett, S. A. (1998). Trait, rater and level effects in 360-degree performance ratings. *Personnel Psychology*, 51, 557–576.
- Mullen, M. R. (1995). Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies*, 26, 573–596.
- Peterson, M. F., & Hunt, J. G. (1997). International perspectives on international leadership. *Leadership Quarterly*, 8(3), 203–231.
- Reilly, R. R., Smither, J. W., & Vasilopoulos, N. L. (1996). A longitudinal study of upward feedback. *Personnel Psychology*, 49, 599–612.
- Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management*, 20(3), 643–671.
- Robie, C., Johnson, K. M., Nilsen, D., & Hazucha, J. F. (2001). The right stuff: Understanding cultural differences in leadership performance. *Journal of Management Development*, 20(7), 639–650.
- Schneider, S. C. (1992). National vs. corporate culture: Implications for human resource management. In V. Pucik, N. M. Tichy, & C. K. Barnett (Eds.), *Globalizing management: Creating and leading the competitive organization* (pp. 159–173). New York: John Wiley & Sons.
- Singh, J. (1995). Measurement issues in cross-national research. *Journal of International Business Studies*, 26, 597–619.
- Smith, C. S., Dugan, S., & Trompenaars, F. (1996). National culture and the values of organizational employees. *Journal of Cross-Cultural Psychology*, 27(2), 231–264.
- Smith, C. S., Tisak, J., Bauman, T., & Green, E. (1991). Psychometric equivalence of a translated circadian rhythm questionnaire: Implications for between- and within-population assessments. *Journal of Applied Psychology*, 76, 628–636.
- Smither, J. W., London, M., Vasilopoulos, N. L., Reilly, R. R., Millsap, R. E., & Salvemini, N. (1995). An examination of the effects of an upward feedback program over time. *Personnel Psychology*, 46, 1–34.
- Spencer-Oatey, H. (1997). Unequal relationships in high and low power distance societies. *Journal of Cross-Cultural Psychology*, 28(3), 284–302.
- Tornow, W. W. (1993). Perceptions or reality: Is multiple-perspective measurement a means or an end? *Human Resource Management*, 32, 221–230.
- Triandis, H. C. (1993). Collectivism and individualism as cultural syndromes. *Cross-Cultural Research: The Journal of Comparative Social Science*, 27, 155–180.
- Van de Vijver, F., & Leung, K. (1997). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology: Theory and method* (Vol. 1, pp. 257–300). Boston: Allyn and Bacon.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4–69.
- Walker, A. G., & Smither, J. W. (1999). A five-year study of upward feedback: What managers do with their results matters. *Personnel Psychology*, 52, 393–423.
- Wohlers, A. J., & London, M. (1989). Ratings of managerial characteristics: Evaluation difficulty, coworker agreement, and self-awareness. *Personnel Psychology*, 42, 235–261.
- Yammarino, F. J., & Atwater, L. E. (1997). Do managers see themselves as others see them? Implications of self-other rating agreement for human resources management. *Organizational Dynamics*, Spring, 25, 35–44.
- Yukl, G. (1998). *Leadership in organizations* (4th ed.). Upper Saddle River, NJ: Prentice Hall.